Tracing the path of forgetting in rule abstraction and exemplar retrieval

Janina A. Hoffmann
janina.hoffmann@uni-konstanz.de
Department of Psychology, University of Konstanz
Universitätsstrasse 10, 78 468 Konstanz, Germany

Bettina von Helversen
b.vonhelversen@psychologie.uzh.ch
Department of Psychology, University of Basel
Department of Psychology, University of Zürich
Binzmühlestrasse 14, 8050 Zürich, Switzerland

Regina A. Weilbächer
regina.weilbaecher@unibas.ch
Department of Psychology, University of Basel
Missionsstrasse 60/62, 4055 Basel, Switzerland

Jörg Rieskamp
joerg.rieskamp@unibas.ch
Department of Psychology, University of Basel

Running title: Forgetting in judgment

Author Note

This research was supported by Swiss National Science Foundation Grant 100014 _146169/1.

Correspondence concerning this article should be addressed to Janina A. Hoffmann, Department of Psychology, University of Konstanz, Universitätsstrasse 10, 78 468 Konstanz, Germany. E-mail: janina.hoffmann@uni-konstanz.de

Abstract

People often forget acquired knowledge over time such as names of former classmates. Which knowledge people can access, however, may modify the judgment process and affect judgment accuracy. Specifically, we hypothesized that judgments based on retrieving past exemplars from long-term memory may be more vulnerable to forgetting than remembering rules that relate the cues to the criterion. Experiment 1 systematically tracked the individual course of forgetting from initial learning to later tests (immediate, one day, and one week) in a linear judgment task facilitating rule-based strategies and a multiplicative judgment task facilitating exemplar-based strategies. Practicing the acquired judgment strategy in repeated tests helped participants to consistently apply the learnt judgment strategy and retain a high judgment accuracy even after a week. Yet, whereas a long retention interval did not affect judgments in the linear task, a long retention interval impaired judgments in the multiplicative task. If practice was restricted as in Experiment 2, judgment accuracy suffered in both tasks. In addition, after a week without practice participants tried to reconstruct their judgments by applying rules in the multiplicative task. These results emphasize that the extent to which decision makers can still retrieve previously learned knowledge limits their ability to make accurate judgments and that the preferred strategies change over time if the opportunity for practice is limited.

Keywords: Judgment, forgetting, rule-based and exemplar-based processes

Tracing the path of forgetting in rule abstraction and exemplar retrieval

One of the earliest discovered laws in psychology is the law of forgetting. The more time has passed between encoding an item and retrieving this item, that is the longer the retention interval is, the less likely people recall the item correctly (Ebbinghaus, 1885; Rubin & Wenzel, 1996). On a class reunion one year after high school, for instance, the names of former classmates may easily come to your mind. Twenty years later, however, you may even encounter problems when naming your former best friends. The course of time makes remembering facts, such as the names of previous classmates (Bahrick, Bahrick, & Wittlinger, 1975), or past events, such as headlines in newspapers (Meeter, Murre, & Janssen, 2005), more difficult.

If people forget information with the passage of time, this should also limit their ability to use this information when making judgments and decisions, affecting judgment quality. Although knowledge about how judgment accuracy varies as time passes by is limited (Ashton, 2000), it seems that not all judgments are equally affected by the time that has passed. For instance, meteorological forecasters have been shown to be more consistent than forecasters in the business or medical domain (Ashton, 2000). This domain difference could be due to people retrieving different information from memory depending on the judgment strategy they rely on.

Suppose, for instance, a hiker tries to forecast every weekend how much rain will fall on a scale from 0 to 40 mm per hour. To judge the precipitation, the hiker may consider how cloudy it is, which shape those clouds have and how strongly the wind blows. If the hiker correctly remembers how important each of those predictors is to forecast the rainfall and applies this policy consistently, her judgment should be independent of the time that passed since her last hike or the number of times she went hiking before.

Alternatively, the hiker may judge the precipitation by remembering how much rain fell on previous hiking tours and how similar the weather conditions are to the weather conditions on previous trips. In this case the judgment will depend on how well the hiker remembers these past hiking tours. Accordingly, if the weather conditions on previous tours are remembered less well the more time has passed, forecasting the rainfall for the current hike should vary depending on the previous hikes that can still be retrieved.

In sum, to predict how judgment accuracy changes over time it is necessary to understand which knowledge people retrieve when making a judgment and how knowledge retrieval changes with the passage of time. So far, however, this question has hardly been studied. The goal of the present research is to fill this gap and to investigate how the passage of time between learning a judgment task and making subsequent judgments influences judgment accuracy and interacts with the way people form their judgments. In the following we will describe the different judgment strategies people may follow and how they may be affected by forgetting in more detail.

Judgment strategies

Evaluating how much rain will fall on a hike requires inferring a continuous criterion, the precipitation, based on a number of features or cues of the judgment object, such as the shape of the clouds and the wind intensity. People learn to solve these judgment tasks by getting feedback about the correct criterion. Judgment research has emphasized the idea that people can base such judgments on two types of judgment strategies: rule-based and exemplar-based (Erickson & Kruschke, 1998; Juslin, Karlsson, & Olsson, 2008; Juslin, Olsson, & Olsson, 2003; von Helversen & Rieskamp, 2008, 2009). These two strategies differ in the way they represent and process knowledge (Hahn & Chater, 1998; Juslin et al., 2003). Rule-based strategies assume that people try to test hypotheses about how each cue relates to the criterion (Brehmer, 1994; Juslin et al., 2008). To judge a new object, people integrate the weighted cue information linear additively (Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Juslin et al., 2003). For instance, dark piled clouds may be judged as predicting more rain. Accordingly, this judgment process requires storing the weight assigned to each cue in long-term memory whereas information about previously encountered objects can be forgotten (Hoffmann, von Helversen, & Rieskamp, 2014; von Helversen & Rieskamp, 2009).

In comparison, exemplar-based judgment strategies assume that people store every previously encountered object, the exemplars, and their associated criterion values in long-term memory (Juslin et al., 2008; Juslin et al., 2003; Nosofsky, 1988). To make a judgment, people retrieve all encountered objects from long-term memory and compare the current object (the probe) to all exemplars. The more similar the probe is to an exemplar, the more this exemplar influences the judgment. Hence, according to exemplar-based judgment strategies the hiker stores each previous hike and the weather conditions in long-term memory. The more the weather conditions match the weather conditions on a previous hike spoiled by rain, the higher will she judge the amount of precipitation.

Research suggests that people can adopt both kinds of judgment strategies, but shift between those strategies depending on the structure of the task (Juslin et al., 2008; Juslin et al., 2003; Karlsson, Juslin, & Olsson, 2007; Pachur & Olsson, 2012; Platzer & Bröder, 2013; von Helversen & Rieskamp, 2009) and characteristics of the decision maker (Hoffmann et al., 2014; Little & McDaniel, 2015; von Helversen, Mata, & Olsson, 2010). Specifically, it has been argued that people are restricted to test linear rules because the comparison processes involved in finding out the importance of each cue are capacity constrained and thus only act upon two successively presented objects (Juslin et al., 2003; Pachur & Olsson, 2012). As a result, on which strategy people rely on should vary with the functional relationship between the cues and the criterion (for a review on effects of functional form across different tasks see Hoffmann, von Helversen, & Rieskamp, 2016; Juslin et al., 2008; Karlsson et al., 2007, 2008). In linear tasks, in which the criterion is a linear, additive function of the cues, people can assess the independent contribution of each cue to the criterion by comparing the difference in attribute values for two judgment objects at a time. In comparison, testing the independent effect of each cue should fail in multiplicative tasks, in which the criterion is a multiplicative function of the cues.

Accordingly, people should deliberately give up testing linear rules and instead memorize single exemplars to solve the judgment task (Karlsson et al., 2008). In line with this idea, it has been found that linear regression models, the predominant account to describe rule-based judgment strategies (Cooksey, 1996; Juslin et al., 2003), capture people's judgments well in linear judgment tasks. In contrast, exemplar models more accurately describe and predict participants' judgments in multiplicative tasks (Hoffmann et al., 2014, 2016; Juslin et al., 2008; Karlsson et al., 2007, 2008; von Helversen & Rieskamp, 2009).

In sum, both rule-based and exemplar-based strategies require to some extent storage in and retrieval from episodic memory, but which kind of knowledge is stored and retrieved varies between the strategies. Whereas rule-based strategies assume that people need to store and retrieve each cue's importance, exemplar-based strategies rely on storage and retrieval of past exemplars. Accordingly, both judgment strategies may be disrupted over time by forgetting, but forgetting may harm rules and exemplars to a different degree.

Sources of forgetting in rule-based and exemplar-based judgments

To what extent people forget information over time is a function of how well the information has been learned initially and if it can be successfully retrieved after some time has passed. The time that passed, however, may not cause forgetting per se, but rather what happened during this time (McGeoch, 1932; Rubin & Wenzel, 1996). Specifically, memory research has postulated two major mechanisms that may cause forgetting: a decay of stored memory traces and interference of similar items (for a historical review see Roediger III, Weinstein, & Agarwal, 2010). Decay theories postulate that memory traces get weaker over time without accessing them. In contrast, interference theories postulate that storing similar items harms retrieval of the to-be-remembered items (Anderson & Neely, 1996; Postman, 1971). Accordingly, other memories compete with the target memory for retrieval and make it more difficult to retrieve the specific target item (Anderson, Bjork, & Bjork, 1994). Which mechanism underlies forgetting over a long time interval is hard to determine, but considering concepts from memory research may inform our understanding about how forgetting may affect retrieval of previously learned rules or exemplars.

Judgment tasks can be thought of as paired-associates learning tasks (Siegel & Siegel, 1972): During learning, people need to form an association between each cue and its importance in rule-based judgments, whereas they need to associate the exemplar (that is, a combination of cues) with its criterion value in exemplar-based judgments. During retrieval, the cues of the presented probe serve as retrieval cues for either the rule or the exemplar.

In rule-based judgments, it has been proposed that people abstract the importance of each cue and adjust (or update) its importance over trials (Hoffmann et al., 2014; Juslin et al., 2008; Pachur & Olsson, 2012). Once a participant has formed a satisfying rule, this rule can be applied to each object. The established rule is hence practiced on every trial. Furthermore, the rule may generalize across different exemplars so that presenting a probe with a different combination of cues interferes with rule retrieval only to a small extent. As a result, rule-based judgments may not be harmed strongly by forgetting. Supporting this idea, Balzer, Rohrbaugh, and Murphy (1983) have found that judgments predicted from a rule-based regression model show a high test-retest reliability even after a week. Actual judgments, however, were less stable over time suggesting that forgetting still intrudes to some degree.

In exemplar-based judgments, it has been proposed that people store each exemplar in a separate memory trace (Estes, 1986). Frequently presented objects are more often encoded facilitating subsequent retrieval of the exemplar. Which exemplar is most similar to the probe, however, varies from trial to trial so that previously stored exemplars are practiced less often and may decay. Furthermore, stored exemplars likely share the same cue value on a particular cue so that the same cue value may activate exemplars with different criterion values and the association between a specific cue and the judgment is more variable. This overlap may increase competition between exemplars, disrupt discrimination between stored exemplars and, in turn, harm retrieval (Capaldi & Neath, 1995). In this vein, it has been found that people follow exemplar-based strategies less, if they cannot discriminate past exemplars from each other (Rouder & Ratcliff, 2004). In sum, exemplar-based judgment strategies may be more prone to forgetting than rule-based judgment strategies.

Previous studies indeed suggest that forgetting may harm retrieval of previously encountered exemplars more than retrieval of previously learned rules. In dot pattern classification paradigms, abstracted prototypes are better remembered over time than single instances (Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Posner & Keele, 1970; Robbins et al., 1978). Furthermore, a recent study applying the looking-at-nothing paradigm found some evidence that people retrieve past exemplars more often in exemplar-based judgments than in rule-based judgments (Scholz, von Helversen, & Rieskamp, 2015). In the looking-at-nothing paradigm (Richardson & Spivey, 2000) participants are presented with objects at different locations. During retrieval, participants tend to look back to the location at which the object they recall was presented suggesting that gaze location indicates which objects people retrieve. Using this looking-at-nothing paradigm, Scholz et al. (2015) found that people who base judgments on similarity look back more often to the location of previously seen similar exemplars than those who base judgments on rules.

If people do not remember previously encountered exemplars well after a long time interval, how can they still solve an exemplar-based judgment task? There is good reason to believe that people try to apply ill-conceived rules if previous exemplars cannot be retrieved (Olsson, Enkvist, & Juslin, 2006). In line with this idea, work relating memory abilities to judgment strategies has found that people with a better episodic memory more frequently adopt an exemplar-based strategy and, in turn, solve exemplar-based judgment tasks more accurately (Hoffmann et al., 2014). Furthermore, people who state that they relied on memory categorize new items more often based on similarity than those who indicated that they learned a rule (Little & McDaniel, 2015). Finally, Bourne, Healy, Kole, and Graham (2006) investigated how participants' stated classification strategy developed over the course of learning and changed after a one-week retention interval in different alphabetical categorization tasks. In the easy and difficult artificial tasks, participants indicated that rule use dominated early in learning, but over the course of learning more memory-based strategies evolved. After a week, however, participants stated that they relearned both tasks by applying a rule and, furthermore, did not revert to the memory-based strategy in the easy task. Accordingly, Bourne et al. (2006) argued that a longer retention interval induces a shift from memory-based to rule-based strategies because rules are better remembered than single instances.

Rationale of the current experiments

Taken together, both rule-based and exemplar-based judgments may involve to some extent storage in and retrieval from long-term memory: In rule-based judgment, people need to retrieve the previously learned rules. In exemplar-based judgment, they need to retrieve previously encountered exemplars and their criterion values. Rules are practiced on every trial and likely generalize across exemplars, whereas previously stored exemplars may be practiced less often and compete for retrieval. Accordingly, prolonging the retention interval between a training and a test phase may harm retrieval of single exemplars more than retrieval of rules.

To manipulate which type of strategy people rely on we varied the functional relationship between the cues and the criterion from a linear to a multiplicative one. In both task structures, participants judge the same objects, but the criterion value associated with each object varies between linear and multiplicative tasks. Linear tasks allow assessing the independent contribution of each cue to the criterion and thus applying linear rules is a viable strategy (Juslin et al., 2008). In comparison, multiplicative tasks require associating a combination of cues with the criterion value, but cannot be solved adequately by testing the independent effect of each cue so that participants should be more likely to rely on exemplar-based strategies (Hoffmann et al., 2014, 2016; Juslin et al., 2008; Karlsson et al., 2007; von Helversen & Rieskamp, 2009). Consequently, we expected a stronger decline in judgment accuracy in multiplicative judgment tasks, which are more likely solved by exemplar memory than in linear judgment tasks in which people should predominantly try to abstract rules.

We tested this prediction in two experiments: Experiment 1 tracks the individual path of forgetting by asking participants to solve either a linear or a multiplicative judgment task and repeatedly retrieve the learned knowledge: immediately, after a day, and after a week. In Experiment 2, we further explore the link between forgetting and judgment accuracy by manipulating the retention interval between participants from immediate recall to recall after a week. Finally, we further tested to what degree forgetting may influence which cognitive strategies people tend to follow at each time point (Bourne et al., 2006).

Experiment 1: Forgetting over time with repeated practice

To test our hypotheses, we trained participants in the present study to predict the criterion value for a number of objects using four cues. In this training session, participants were randomly assigned to one of two judgment tasks: a linear judgment task to induce a rule-based judgment strategy or a multiplicative judgment task to induce an exemplar-based judgment strategy. To induce forgetting, we asked participants to judge old items (objects encountered in training) as well as new items (unknown objects) repeatedly at three retention intervals: an immediate test session, a test session after one day, and a test session after one week.

Method

Participants. 83 participants (53 female, 30 male, $M_{\text{Age}} = 24.6$, $SD_{\text{Age}} = 6.5$) were recruited at the University of Basel and randomly assigned to the linear ($n = 41$) or the multiplicative task ($n = 42$). Two participants who did not show up for all sessions were excluded from the study (one participant in the linear, and one in the multiplicative task) as well as one who was assigned to the wrong task in one of the sessions. Participants received course credit or 20 Swiss Francs (CHF) per hour for participating in the experiment. In addition, they could earn a bonus based on their judgment performance (M = 6.06 CHF, SD = 2.11 CHF). The first session took about an hour, whereas the second and the third lasted approximately 30 minutes.

Design and material. The cover story asked participants to predict how long the pupal stage lasts for different fictitious butterfly species on a scale from 10 to 20 days. The butterflies' appearance differed in four binary features (the cues): wing color (red vs. violet), antennae color (black vs. orange), color of stripes (brown or pink), and shape of spots (oval or serrated). Figure 1 shows two sample butterfly species with different cue values on all cues. These cues could be used to predict how long the pupal stage for a butterfly lasts (the criterion). In the linear judgment task, the criterion was a linear, additive function of the cues,

$$y_{\text{lin}} = 4x_1 + 3x_2 + 2x_3 + x_4 + 10. \tag{1}$$

Each cue, x_1, x_2, x_3, and x_4, could take a cue value of zero or one. The cue weights were randomly assigned to the four pictorial cues, as were the cue values (zero or one) to the features (e.g., oval or serrated spots). In the multiplicative judgment task, the criterion was a nonlinear, multiplicative function of the cues:

$$y_{\text{mult}} = 9 + e^{(4x_1+3x_2+2x_3+x_4)/4.15} \tag{2}$$

Table 1 illustrates the task structure: The cues were given a binary value of zero or one, and they varied in their cue weights; that is, in their importance for predicting the criterion. In principle, the rule-based model can perfectly solve a linear task, but approaches multiplicative tasks less well because it is restricted to linear, additive rules.

The exemplar model can learn to solve both types of judgment tasks perfectly, if training exemplars are repeated. Considering as well that participants often do not reach perfect accuracy in these judgment tasks, it is thus difficult to distinguish the exemplar model from a rule-based model based on trained exemplars. Therefore we introduced new, unseen objects to allow the models' to make different predictions and to test them rigorously against each other. From all possible items, we selected a training set of 10 old items and a test set of 6 new items so that the rule-based model (with 5 parameters, one cue weight for each cue and the intercept) and the exemplar model (with one sensitivity parameter and equal attention weights) made different predictions for new items in both judgment tasks (see Appendix A for model descriptions). As illustrated by the models' predictions, both models made accurate predictions for old items in the linear judgment task (measured in root mean square deviations between model predictions and the correct criterion, RMSD = 0), but the rule-based model (RMSD = 0) predicted the criterion values of new items more accurately than the exemplar model (RMSD = 1.93). In the multiplicative task, an exemplar model better fitted the old items (RMSD = 0) than a rule-based model (RMSD = 0.95) and made slightly better predictions for new items (RMSD = 1.86) than the rule-based model (RMSD = 2.19). In comparison, a guessing model, a model predicting the mean of the criterion values in every trial, results in a RMSD = 3.1 at the end of training in the linear task and an RMSD = 3.2 in the multiplicative task. Exact predictions of each model depend on the estimated parameters for each participant.

One potential problem with manipulating the strategy with different task structures is that the task structure could also influence the ability to remember the training exemplars. If the multiplicative task structure we chose led per se to a higher rate of forgetting than the linear one, an exemplar-based learning model, ALCOVE (Kruschke, 1992), should predict a higher judgment error for old items in the multiplicative task than in the linear one. We modelled forgetting by assuming interference such that over time more and more new exemplars would be encountered that then would interfere with retrieving the previously learnt ones (for details see Appendix B). ALCOVE predicted that forgetting of exemplars would cause a similar increase in judgment error for old items in the linear and multiplicative task we used, independent of the degree of interference. Repeating this simulation with parameters sampled from the best fitting parameters estimated from participants' training data similarly led to the conclusion that ALCOVE predicts the same rate of forgetting for the linear and the multiplicative task we used. This suggests that if judgment accuracy changes over time to a different degree in the linear and multiplicative task, this differential forgetting cannot be explained by how easily an exemplar-based strategy forgets the items in the two tasks.

Procedure. The judgment task consisted of a training session and three test sessions. In the training session, participants learned to estimate the criterion values for the 10 old items from the training set. In each trial, participants first saw a picture of a butterfly and were asked to estimate its criterion value. Afterwards they received feedback about the correct value, their own estimate, and the points they had earned. The training session ended when a learning criterion was reached. Participants met this learning criterion when judgment accuracy, as measured in root-mean-square deviation (RMSD) between participants' judgments and the criterion values in one training block, fell below 1 RMSD. We employed this learning criterion to minimize the possibility that differential forgetting in judgment could have resulted from initial differences in judgment accuracy between tasks and to achieve a high judgment accuracy at the end of training. Each participant completed at least 10 training blocks, each consisting of the 10 old items; training terminated after 20 blocks even if the learning criterion had not been reached. Earlier work has set a more lenient learning criterion of 1.5 RMSD to be met within fewer training blocks (Mata, von Helversen, Karlsson, & Cüpper, 2012; von Helversen et al., 2010) suggesting that participants may meet our learning criterion as well within 20 training blocks. In the test sessions, participants estimated criterion values for all 16 butterflies, 10 old and 6 new ones, six times without getting any feedback. Presentation order in each training and test block was randomly determined.

To motivate participants to reach a high performance, participants could earn points in every trial. Participants earned 10 points for a correct answer and 5 points if their judgment deviated by 1 from the correct answer. At the end of the judgment tasks, the points earned were converted to a monetary bonus (500 points = 1 CHF). In addition, participants earned a bonus of 5 CHF if they reached the learning criterion for the judgment task within 20 training blocks. Participants returned to the lab after 24 h as well as after one week to repeat the test session of the judgment task.

Results

Learning success at the end of training. Overall, the number of participants reaching the learning criterion varied slightly between the judgment tasks, but the difference was not significant, $\chi^2(1) = 2.20$, $p = .138$. In the linear judgment task, 25 out of 40 participants reached the learning criterion (62.5%), whereas in the multiplicative task 32 out of 40 participants (80%) mastered the training phase successfully. Among those participants who did not learn the task, three participants in the multiplicative task and four participants in the linear task did not outperform a guessing model. Descriptively, participants needed slightly more training blocks in the linear than in the multiplicative task, $t(78) = 1.6$, $p = .108$ (see Table 2 for descriptive statistics). The number of training blocks participants needed was highly correlated with judgment error in both tasks (linear: $r = .75$; multiplicative: $r = .78$). Taken together, participants learned the judgment tasks on average equally well suggesting that the multiplicative task was not more difficult than the linear one.

Judgment performance over time. According to our hypothesis, increasing the retention interval between training and test should increase judgment error more in the multiplicative than in the linear judgment task. This increase in judgment error should be most pronounced for old items because people should be more likely to forget specific training exemplars than previously learned rules. To compare judgment performance for old items across time, we measured judgment error in the training session as the RMSD between participants' judgments in the last training block and the correct criterion. Judgment error in the three test sessions (immediate test, test after 1 day, and after 1 week) was measured as the RMSD between the criterion and participant's judgments, averaged over the six presentations in each test session. Figure 2 shows judgment error for old and new items in each test session separately for the linear (white bars) and the multiplicative judgment task (gray bars). Descriptively, participants achieved a similar accuracy level for old items at the end of training in the linear and the multiplicative judgment task. Judgment error in the linear task is equally high (even slightly lower) as in studies using a similar design (Mata et al., 2012; von Helversen et al., 2010). In the linear judgment task, judgment error remained constant across time from immediate test to the next day to one week (d = -0.07 from last block of training to one week using an effect size based on the change score for repeated measures, Morris & DeShon, 2002). In the multiplicative task, however, judgment error rose for old items from immediate test to the next day to a week later (d = 0.70 from last block of training to one week).

To test the hypothesis that a longer retention interval harms exemplar-based judgments more than rule-based judgments on old training items we conducted a repeated-measures ANOVA using judgment error for old items as the dependent variable; judgment task (linear vs. multiplicative) was included as the between-factor and session (training, immediate test, test after 1 day, and test after 1 week) as the within-factor. Tests for the within-factor were corrected for sphericity using the Greenhouse-Geisser method. The type of judgment task did not affect judgment error, $F(1, 78) = 0.4$, $\eta^2 = .004$, $p = .531$, but participants made less accurate judgments in later test sessions, $F(2.4, 184.4) = 5.9$, $\eta^2 = .009$, $p = .002$. Furthermore, session interacted with the type of judgment task, $F(2.4, 184.4) = 7.7$, $\eta^2 = .01$, $p < .001$, suggesting that judgment error increased more over time in the multiplicative than in the linear judgment task. To further investigate in which sessions judgment error increased the most, we set Bonferroni-corrected contrasts on the marginal means for judgment error in the different sessions. Specifically, we compared judgment error in one session to the average error of the subsequent sessions to identify between which sessions judgment error increased. In the multiplicative task, comparing judgment error at the end of training to average judgment error in the three test sessions suggested that participants made more errors in the test sessions than at the end of training, $\Delta\hat{x} = 0.41$ ($\Delta\hat{x}$ reflects the difference in estimated least square means), SE = 0.07, $t(234) = 5.6$, $p < .001$. Judgment error further increased from immediate test to the two delayed tests, $\Delta\hat{x} = 0.21$, $SE = 0.08$, $t(234) = 2.8$, $p = .038$, but did not change from the delayed test after a day to test after one week, $\Delta\hat{x} = 0.10$, $SE = 0.09$, $t(234) = 1.10$, $p = 1.00$. In the linear judgment task, judgment error increased neither from the last block of training to the three test sessions, nor from immediate test to delayed tests, nor from one day to one week (all $\Delta\hat{x} < |0.06|$, all $p = 1.00$). [1]

With regard to new items, judgment error descriptively increased in the multiplicative judgment task from immediate test to delayed test after a day and test after a week ($d = 0.26$ from immediate to one week). Likewise, judgment error increased for new items in the linear judgment task from immediate test to test after one day to test after one week ($d = 0.33$). Reflecting the descriptive results, a repeated measures ANOVA (Greenhouse-Geisser corrected) indicated that participants made worse judgments on new items in the multiplicative judgment task than in the linear judgment task, $F(1, 78) = 14.9$, $\eta^2 = .15$, $p < .001$, and judgment error rose in later test sessions, $F(1.6, 128.3) = 5.1$, $\eta^2 = .01$, $p = .01$, but the type of judgment task did not affect how strongly judgment error increased over the test sessions, $F(1.6, 128.3) = 0.2$, $\eta^2 < .01$, $p = .745$.

[1] Excluding participants based on the learning criterion slightly changed results suggesting in addition that participants in the linear task performed better across all sessions, $F(1, 55) = 6.7$, $\eta^2 = .08$, $p = .012$, but judgment error still increased across sessions, $F(2.4, 130.0) = 12.9$, $\eta^2 = .06$, $p < .001$, and increased more strongly in the multiplicative than in the linear task, $F(2.4, 130.0) = 5.0$, $\eta^2 = .03$, $p = .006$. The pattern for strategy use remained the same so that we report results for all participants.

Taken together, we found that prolonging the retention interval between training and test increased judgment errors on old training items in the multiplicative judgment task, but not in the linear judgment task. If participants had to generalize their judgment to new items, we found that participants in the multiplicative judgment task made on average more errors than participants in the linear judgment task. A longer retention interval, however, did not increase judgment error on new items more in the multiplicative than in the linear task.

Judgment strategies over time. Previous research (Bourne et al., 2006) has suggested that a long retention interval leads to a shift from exemplar-based strategies to rule-based strategies because people cannot retrieve previously encountered exemplars and instead they relearn the task by applying a rule. According to this hypothesis, participants in the multiplicative judgment task should shift from exemplar-based strategies in the immediate test to rule-based strategies after a week. However, in contrast to Bourne et al. (2006), in our study participants had the possibility to repeatedly practice their judgment strategy both in the immediate test and after a day, making it likely that they did not need to abandon an exemplar-based judgment strategy after a week. To determine how much support is provided for the exemplar strategy over the rule-based strategy, we relied on a cognitive modeling approach. We fitted a linear regression model serving as a rule-based strategy and an exemplar model with four attention weights to participants' judgments, separately for each of the three test sessions and each participant (see Appendix A for more details on modelling and Table A2 for participants' strategy classifications). To account for random guessing, we compared those models to a baseline model (a model estimating participants' mean judgment). All models were estimated representing judgment errors with a beta distribution.[2] For a few participants the evidence favored most strongly the baseline model in one session or more (linear: n = 3, multiplicative: n = 3) and so we did not consider those participants further in subsequent analyses.

[2] We thank an anonymous reviewer for proposing a change in the error distribution.

Figure 3 illustrates how well the model predictions (diamonds) of the rule-based and the exemplar model match participants' responses (crosses) for those participants unequivocally classified to the rule-based model (first and third row) and the exemplar model (second and fourth row). The upper two rows illustrate model predictions and participants' judgments in the linear task in immediate test, test after one day and test after one week; the lower two rows illustrate model predictions and participants' judgments in the multiplicative task across time. In the linear task, judgments of participants best described by the rule-based model on average match criterion values and model predictions well, whereas in line with the predictions of the exemplar model judgments deviate more from the criterion values for participants best described by the exemplar model. In the multiplicative task, participants best described by the rule-based model more likely overestimated low criterion values and underestimated high criterion values, whereas participants best described by the exemplar model judged the criterion values on average more accurately.

We then calculated the evidence ratio between the exemplar strategy over the rule-based strategy (Wagenmakers & Farrell, 2004) that expresses as a normalized probability how much support is provided for the exemplar strategy over the rule-based strategy ranging from 0 (evidence in favor of the rule-based strategy) to 1 (evidence in favor of the exemplar strategy). In all test sessions, this evidence ratio was descriptively higher in the multiplicative task than in the linear judgment task (see table 2) suggesting that more participants were better described by an exemplar model in the multiplicative compared to the linear task. Thus, in the linear task linear rules best described participants' judgments, whereas in the multiplicative task a substantial proportion of participants seems to have preferred an exemplar-based strategy although a substantial number of participants was still well-described by rules.

To assess how closely the judgment strategies in immediate test corresponded to the judgment strategies after a day or after a week, we calculated Spearman's rank correlations between the evidence ratios. Spearman's r was used because it does not presume a linear relationship, an assumption that may not be appropriate for evidence ratios close to the bounds. In the multiplicative task, stronger evidence for the exemplar strategy in immediate test was associated with stronger evidence for the exemplar strategy after a day, Spearman's $r = 0.75$, but the relationship was slightly less pronounced after a week, Spearman's $r = 0.60$. In the linear task, stronger evidence for the exemplar model was similarly associated with stronger evidence for the exemplar strategy after a day, Spearman's $r = 0.74$, but slightly lower a week later, Spearman's $r = 0.50$.

To investigate to what extent participants shifted between strategies as a function of task and retention interval, we conducted a beta regression using the evidence ratio as the dependent variable (Smithson & Verkuilen, 2006). Using a beta regression as a statistical model is necessary because the evidence ratios are bound to scale between 0 and 1 with most values approximating the upper or lower end of the scale (see Appendix A for a short introduction). In nested models, effects of the independent variables, here for instance task and retention interval, can be tested via model comparison using Likelihood ratio tests or Akaike's Information Criterion (AIC; Smithson & Verkuilen, 2006). The log-likelihood ratio test compares the full model against a more restricted version of the full model. The AIC can be used to compare as well non-nested model and penalizes more complex model by the number of model parameters. Models with smaller AIC values are preferred over models with higher AIC values.

Overall, including judgment task as a predictor in the beta regression improved model fit (AIC = -945) compared to a baseline model estimating only the intercept, AIC = -938, $\chi^2(1) = 8.3$, $p = .004$. This main effect indicated that participants in the multiplicative task had a higher chance than participants in the linear task that the exemplar model outperformed the rule-based model, $OR = 1.7$, CI = [1.2; 2.4]. Adding the test session as a predictor did not further improve the fit of the model, AIC = -944, $\chi^2(2) = 3.2$, $p = .200$, nor did accounting for a possible interaction, AIC = -940, $\chi^2(2) = 0.1$, $p =$

.949. In sum, the type of judgment task predicted which judgment strategy described participants' judgments best at each time point, but a longer retention interval did not increase the number of participants best described by rules suggesting that participants did not shift to rule-based judgment strategies in response to a longer retention interval.

Discussion

In Experiment 1, we investigated whether prolonging the retention interval affects how accurately people make judgments in two different kinds of judgment tasks: a linear judgment task that can best be solved by abstracting linear, additive rules and a multiplicative judgment task that can better be approached by storing and retrieving exemplars from long-term memory. In line with our hypothesis, we found that judgment accuracy for old items encountered in training dropped more from training to recall after a week in the multiplicative than in the linear judgment task. In the linear judgment task, participants judged —on average —old items as accurately after a week as at the end of training, whereas judgment errors increased from the last training block to test after a week in the multiplicative judgment task. This result matches previous research suggesting that people remember abstracted knowledge, for instance in the form of prototypes, better than single instances after a long retention interval (Homa et al., 1973; Posner & Keele, 1970; Robbins et al., 1978) and supports the idea that exemplar-based judgments build to a stronger extent on episodic memory than rule-based judgments (Hoffmann et al., 2014).

Replicating previous work on judgment strategies (Hoffmann et al., 2014; Juslin et al., 2008), we found that more participants were best described by an exemplar model in the multiplicative task than in the linear task. With regard to the question of how judgment strategies developed over time, our results suggest that participants relied consistently on the same judgment strategy across time: In the linear task, most participants were still best described by a rule-based strategy after a week. Similarly, an equal number of participants were best described by the exemplar model in immediate test and after one week. This result suggests that participants still tried to retrieve previously encountered exemplars after one week. This finding differs from previous research (Bourne et al., 2006) suggesting that people prefer relearning complex categorizations by relying on rules, although they stated that they previously solved the task by retrieving exemplars from memory.

One reason why participants potentially did not shift from an exemplar-based strategy to a rule-based strategy is that they had the possibility to repeatedly practice their judgment strategy. Practicing a task even without getting feedback can benefit long-term retention in a wide range of tasks from free recall to function learning and may outperform studying the correct solution (Kang, McDaniel, & Pashler, 2011; Karpicke & Roediger III, 2008). One explanation why practice is so beneficial for retention focuses on the idea that those repeated retrieval processes may strengthen the memory trace by elaboration, deeper encoding or adding multiple retrieval routes (Roediger & Karpicke, 2006). It is possible that asking participants to solve the judgment task immediately, after one day, *and* after one week, involved such repeated retrieval processes. Therefore, our design that tried to track individual paths of forgetting might have prevented a high amount of forgetting in the judgment task. To circumvent the possibility that repeated practice may have restricted the amount of forgetting, we tested in a second experiment whether forgetting impacts judgments more and participants shift to a greater extent to rule-based strategies if they do not have the opportunity to repeatedly practice the judgment task at several time points.

Experiment 2: Forgetting over time without repeated practice

In Experiment 2, we studied how forgetting affected participants' judgments if they did not have the possibility to practice their judgment strategy between training and a later test. As in Experiment 1, participants learned to solve either a linear or a multiplicative judgment task in a training session. To induce forgetting, we asked participants to rejudge these old items as well as new items either immediately after training or after one week.

In addition, we assessed recognition memory for old items in a two-alternative forced-choice test. Past research has found that participants who possess a better episodic memory are more likely to adopt an exemplar-based strategy and, in turn, make more accurate judgments in multiplicative tasks (Hoffmann et al., 2014). This suggests that people using an exemplar strategy may remember better which objects they encountered during training than rule users. In contrast, if participants in the linear task only learned to abstract rules, they should not be able to discriminate old from new exemplars. However, there is some research suggesting that the relationship between recognition and strategy use is more complex. First, if people are asked to recognize the previously encountered exemplars in a recognition test they are better at remembering items violating a salient knowledge structure, for instance a rule, than items following the knowledge structure (Davis, Love, & Preston, 2012; Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). Second, the more salient the rules are in these rule-plus-exception tasks, the better the exceptions violating the rule are remembered (Sakamoto & Love, 2004). This result could indicate that also rule users can perform well in a multiplicative task if they can remember the exceptions to the rule well. Lastly, it has been found in rule-plus-exception tasks that previously encountered items consistent with a rule show a recognition advantage over novel items implying that people possess some residual memory for old exemplars, although they abstracted a rule (Palmeri & Nosofsky, 1995). Accordingly, it is possible that learners in the linear judgment task encode both a rule-based and an exemplar-based representation and therefore show as well a recognition advantage for previously encountered old items over new items. This recognition advantage may then reduce differences in recognition performance between strategies and tasks.

Method

Participants. 142 participants (115 female, 27 male, $M_{\text{Age}} = 24.3$, $SD_{\text{Age}} = 6.4$) were recruited at the University of Basel. Participants were randomly assigned to one of the four conditions: 35 to the linear task with immediate recall, 37 to the linear task with recall after a week, 33 to the multiplicative task with immediate recall and 37 to the multiplicative task with recall after a week. Two participants who did not show up for the test session after one week were excluded from the study (one participant in the linear, and one in the multiplicative task) as were three participants who were assigned to the wrong condition. Participants received course credit or 20 Swiss Francs (CHF) per hour for participating in the experiment. In addition, they could earn a performance-dependent bonus (M = 5.17 CHF, SD = 2.68 CHF). The training session took about an hour, whereas the test session took approximately thirty minutes.

Procedure. Material and procedure followed closely the experimental set-up of Experiment 1. Participants were randomly assigned to the linear or the multiplicative judgment task. In contrast to Experiment 1, we varied the retention interval between training and test between participants: Half of the participants in each judgment task solved the test session immediately after training whereas the other half returned to the lab after a week.

After participants completed the test session, they solved a two-alternative forced-choice recognition test. In each trial, participants saw one "old" butterfly —that is, one they already knew from training —and one "new" butterfly —that is, a butterfly from the test set introduced in the test session. Participants had to determine which of those two butterflies was "old"; that is, the one they already knew from training. All 10 old butterflies were presented twice with each of the 6 new butterflies, resulting in 120 forced choice decisions.

Results

Learning success at the end of training. As in Experiment 1, the number of participants reaching the learning criterion did not vary strongly between the judgment tasks and the retention intervals, $\chi^2(4) = 3.7$, $p = .448$. In the linear judgment task, 20 out of 35 participants (57.1%) assigned to immediate test reached the learning criterion as did

19 out of 36 participants (52.8%) assigned to test after a week. The multiplicative task was mastered successfully by 23 out of 31 participants (74.2%) assigned to immediate test and by 21 out of 35 participants (60.0%) assigned to the test one week later. Among those participants who did not learn the task, five participants in the linear task (immediate: 2, one week later: 3) and 12 participants in the multiplicative task (immediate: 4, one week later: 8) did not outperform a random guessing model. Table 3 displays descriptive statistics for both judgment tasks, separately for immediate test and test after one week. In the multiplicative task, participants needed —on average —slightly fewer training blocks than participants in the linear judgment task, but this difference did not reach significance, $t(135) = 1.8$, $p = .071$. As in Experiment 1, judgment error in the last training block and the number of training blocks needed were highly correlated ranging from $r = .69$ in the linear task for test after a week to $r = .77$ in the linear task for immediate test. In sum, these results suggest that participants did not solve the linear task more easily than the multiplicative task.

Judgment performance over time. As in Experiment 1, we expected a longer retention interval to impede judgment accuracy most severely on old items in the multiplicative judgment task. Figure 4 illustrates judgment error on old and new items for the last block of training and the test session, separately for the judgment tasks and retention intervals (see Table 3 for descriptive statistics). In the linear judgment task, participants who took the immediate test were descriptively as accurate on old items in test as in the last block of training (d = -0.19, d based on the change score for repeated measures, Morris & DeShon, 2002, $\Delta\hat{x}$ = -0.09, $SE = 0.14$, $\Delta\hat{x}$ for difference in marginal means), whereas participants who solved the test session a week later made more errors on old items in test than in the last block of training ($d = 0.21$, $\Delta\hat{x} = 0.21$, $SE = 0.14$). In the multiplicative judgment task, participants who took the immediate test made only slightly more errors on old items in test than in the last block of training ($d = 0.24$, $\Delta\hat{x} = 0.15$, $SE = 0.15$), whereas participants who solved the test session a week later made worse judgments on old items in test than in the last block of training (d = 0.64, $\Delta\hat{x}$ = 0.74, SE = 0.14). Finally, participants who solved the linear judgment task a week later made on average fewer errors than participants who solved the multiplicative task, d = -0.57.

To test the hypothesis that a longer retention interval harms exemplar-based judgments more than rule-based judgments for old items we conducted a repeated-measures ANOVA on judgment error using retention interval, judgment task, and session (training vs. test) as independent variables. Overall, participants made fewer errors in the last training block than in test, $F(1, 133) = 12.2$, $\eta^2 = .01$, $p < .001$, but the judgment task did not affect judgment errors, $F(1, 133) = 2.1$, $\eta^2 = .01$, $p = .150$. A longer retention interval increased judgment error, $F(1, 133) = 6.9$, $\eta^2 = .04$, $p = .010$. An interaction between retention interval and session indicated that judgment error increased more strongly between training and test for those participants who took the test after a week than immediately, $F(1, 133) = 9.4$, $\eta^2 = .01$, $p = .003$. Further, an interaction between judgment task and session indicated that judgment error increases more from training to test for participants in the multiplicative task than for participants in the linear task, $F(1, 133) = 7.1$, $\eta^2 = .01$, $p = .009$. Yet, in contrast to our hypothesis that a longer retention interval contributes to more errors in the multiplicative than in the linear judgment task, neither the interaction between retention interval and judgment task, $F(1, 133) = 0.5$, $\eta^2 = .003$, $p = .476$, nor the three-way interaction was significant, $F(1, 133) = 1.0$, $\eta^2 = .001$, $p = .320$.[3]

With regard to new items, participants in the linear task descriptively made more errors if they were tested a week later than if they took an immediate test (d = 0.10). Similarly, in the multiplicative task participants who were tested after a week made less accurate judgments than those who took an immediate test (d = 0.41). Furthermore, participants made less accurate judgments in the multiplicative task than in the linear task both in immediate test (d = 0.57) and after a week (d = 0.88). To investigate how a longer

[3] Excluding participants based on the learning criterion did not change results for judgment accuracy on old items nor for strategy classifications.

retention interval affected judgment errors for new items we conducted an ANOVA on judgment error using retention interval and judgment task as independent factors. Judgments were less accurate in the multiplicative compared to the linear task, $F(1, 133) = 18.0$, $\eta^2 = .12$, $p < .001$, but neither retention interval, $F(1, 133) = 2.1$, $\eta^2 = .02$, $p = .150$, nor its interaction with the type of task affected judgment accuracy, $F(1, 133) = 0.7$, $\eta^2 < .01$, $p = .412$.

In sum, a longer retention interval impeded judgment accuracy on old items in both judgment tasks. Judgment error increased more from training to test for those participants who took a delayed test after a week than for those who took an immediate test. Furthermore, participants in the multiplicative judgment task were less successful at generalizing their performance to new items than participants in the linear task, independent of the retention interval.

Judgment strategies over time. In Experiment 2, participants did not have the possibility to practice their judgment strategy between training and delayed test. Without practicing the judgment strategy, it is possible that participants shift from an exemplar-based judgment strategy to a rule-based judgment strategy after a week (Bourne et al., 2006). To describe judgment strategies, we fitted an exemplar model, a rule-based model and a baseline model to participants' judgments in each test session. As in Experiment 1, we excluded for all subsequent analyses those participants for whom the evidence favoured a baseline model in the linear (immediate: $n = 1$; one week: $n = 2$) and the multiplicative task (immediate: $n = 3$; one week: $n = 4$). For all remaining participants, the evidence ratio favoured the rule-based model in the linear task (see table 3). In the multiplicative task, the evidence ratio supported more strongly the exemplar model in immediate test, but provided more support for the rule-based model in test after a week.

Figure 5 illustrates —analogously to Figure 3 —model predictions as well as participants' average responses for participants unambiguously classified to the rule and the exemplar model. The upper two rows illustrate model predictions and participants' judgments in the linear task in immediate test and test after one week; the lower two rows illustrate model predictions and participants' judgments in the multiplicative task across time. In the linear task, model predictions and participants' responses for the rule-based model more closely match the criterion values, whereas model predictions and participants' responses showed a higher variability for the exemplar model. In the multiplicative task, judgments of participants classified to the exemplar model were better calibrated in immediate test, whereas participants and model predictions for the rule-based model suggested an overestimation of smaller criterion values. Furthermore, the plots highlight that after one week participants' judgments as well as the model predictions on average match the criterion values less well than in immediate test.

To understand how strategies may change depending on the type of judgment task and retention interval, we conducted a beta regression using the evidence ratio as the dependent variable and judgment task as well as retention interval as predictors. Overall, this analysis suggested that including judgment task as a predictor improved model fit (AIC = -409) compared to a baseline model estimating only the intercept, AIC = -406, $\chi^2(1) = 4.7$, $p = .03$. Furthermore, adding retention interval as a predictor suggested a main effect of retention interval, AIC = -417, $\chi^2(1) = 9.7$, $p = .002$. Yet, these main effects were qualified by an interaction between judgment task and retention interval, AIC = -422, $\chi^2(1) = 7.6$, $p = .006$. The interaction model suggested that the evidence ratio favoured the exemplar model in the multiplicative task compared to the linear task, $OR = 3.7$, CI = [1.9; 7.3] and that across both tasks the retention interval did not change the evidence for the exemplar model, $OR = 0.83$, CI = [0.46; 1.51]. Yet, in the multiplicative task evidence for the exemplar model was reduced a week later, $OR = 0.28$, CI = [0.11; 0.68]. Taken together, those results indicate that participants may have shifted from a memory-based exemplar strategy in immediate test to a rule-based strategy after a week.

Predicting recognition memory with judgment strategies. Finally, we assessed in both tasks to what extent strategy use can predict recognition memory for previously encountered exemplars across time. On the one hand, participants relying on an exemplar-based strategy may rely more on episodic memory and discriminate old from new items better than participants relying on rules. On the other hand, it is possible that also rule-based learners may possess some residual memory for old exemplars (Palmeri & Nosofsky, 1995) and are likewise able to discriminate old from new items. Overall, participants correctly recognized 63.0% (*recognition rate*, SD = 18 %) of all old items. In the linear judgment task, the recognition rate was higher in test after a week than in immediate test (see Table 3). In the multiplicative task, participants recognized the old items slightly worse after a week than in immediate test.

To investigate how judgment strategies and retention interval affected recognition memory, we conducted a logistic regression using the number of correctly and incorrectly recognized old items as dependent variable and predicted this success rate with retention interval, judgment task, and judgment strategy, as measured with the evidence ratio. Tests of parameter estimates were conducted with a Likelihood ratio test. Overall, this logistic regression model described the data well, AIC = 2796 (compared to AIC = 2931 for the null model), Nagelkerke's $R^2 = 0.69$, with on average a higher recognition rate in the linear than in the multiplicative judgment task, $\chi^2(1) = 36.7$, $p < .001$. Furthermore, this analysis suggested a three-way interaction between judgment task, retention interval, and judgment strategy, $\chi^2(1) = 15.0$, $p < .001$.

Therefore, we broke up this interaction by separately analyzing the judgment tasks. In the linear task, participants had a higher recognition rate after a week than in immediate test, $\chi^2(1) = 45.7$, $p < .001$, $OR = 1.37$, CI = [1.23, 1.52]. Furthermore, participants who were more likely better described by the rule-based model recognized more old items correctly than participants more likely better described by the exemplar model, $\chi^2(1) = 15.5$, $p < .001$, $OR = 0.81$, CI = [0.68, 0.95]. The recognition rate at different retention intervals did not vary with the evidence provided for each model, $\chi^2(1) = 0.8$, $p = .374$, $OR = 0.89$, CI = [0.68, 1.16].

In the multiplicative task, Likelihood ratio tests suggested that participants overall had a lower recognition rate after a week than in immediate test, $OR = 0.84$, CI = [0.72, 0.98], $\chi^2(1) = 6.7$, $p = .010$, and the recognition rate varied with the evidence for the exemplar model $\chi^2(1) = 22.6$, $p < .001$. Yet, these main effects were qualified by an interaction, $\chi^2(1) = 20.9$, $p < .001$. This interaction suggested that the evidence for the exemplar model did not change recognition rate per se, $OR = 1.08$, CI = [0.92, 1.28], but after a week stronger evidence for the exemplar model led to a higher recognition rate, $OR = 1.89$, CI = [1.44, 2.49].

Figure 6 shows the recognition rate for each old item, plotted over participants' average judgment for this item considering only those participants who are unequivocally classified to one strategy. The graph illustrates that participants who are classified to the rule-based model in the linear task recognize old items better than participants who are classified to the exemplar model. In the multiplicative task, after a week participants classified to the exemplar model recognize old items better than participants who are classified to the less suitable strategy. However, exemplar users also show larger standard errors than rule-users after a week indicating that recognition memory has a higher variability. The reason for this finding is possibly that only a few participants still adopt an exemplar-based strategy after one week.

Discussion

Instead of tracking the individual course of forgetting, Experiment 2 varied the retention interval between participants to reduce the possibility that repeated practice of judgment strategies limited the decline of judgment accuracy over time. In line with the results from Experiment 1, we found that participants in the multiplicative task judged old items less accurately in immediate test than in training, whereas participants in the linear judgment task kept their performance in immediate test. In contrast to our hypothesis, however, judgment error increased more strongly in a delayed test after a week not only in the multiplicative task, but also in the linear judgment task. Accordingly, a longer retention interval harmed judgments both in the multiplicative and in the linear judgment task. Furthermore, in contrast to Experiment 1, judgment strategies were not stable over time, but changed across time: In immediate test, participants were likely better described by the rule-based model in the linear task, whereas the evidence preferred an exemplar model in the multiplicative task. After a week, however, participants relied less on an exemplar model in both judgment tasks suggesting that if participants do not have the opportunity to practice an exemplar-based judgment process, they shift to a greater extent to rule-based strategies.

General discussion

The passage of time makes it harder to remember previously learned knowledge (Ebbinghaus, 1885; Rubin & Wenzel, 1996), but it can also impede previously acquired skills, such as speaking foreign languages (Bahrick, 1984). Although forgetting affects a wide range of cognitive abilities, only a few studies in judgment research have paid attention to such basic memory phenomena. Our research tried to shed light on the question of how a longer retention interval may change the knowledge people retrieve to make a judgment and, ultimately, judgment accuracy. Reinterpreting judgment tasks as paired-associates learning tasks, we argued that people may need to form different associations when learning to solve rule-based and exemplar-based judgment tasks: In rule-based judgments, people should associate each cue with its importance, whereas they need to associate exemplars with their corresponding criterion value in exemplar-based judgments. In a later test phase, people retrieve either previously learned rules or exemplars. Specifically, we hypothesized that storing a range of similar exemplars may make exemplar-based judgments highly vulnerable to forgetting, whereas rules receive more training, are likely generalized to a range of different objects, and may hence be forgotten less easily. We tested this hypothesis in two experiments: In a first experiment that tracked judgment performance over a week we found that judgment error on old items increased more within this week in the multiplicative than the linear judgment task, reflecting the idea that forgetting over time harms successful retrieval of single exemplars more than retrieval of rules. Varying the retention interval between groups in a second experiment, we found that judgment performance on old items decreases not only in multiplicative tasks, but also in linear ones indicating that previously learnt rules can also be forgotten.

Looking more closely at the temporal curve of forgetting, we found that judgment error for previously encountered items in the multiplicative task already increased between the end of training and immediate test in both experiments, whereas participants in the linear judgment task kept their performance from training to immediate test. Possibly, introducing new items in the multiplicative task already interferes with retrieving old training exemplars so that participants likely confuse old training items with novel items. Yet, our results provide mixed evidence for the idea that prolonging the retention interval leads to a greater amount of forgetting in exemplar-based than in rule-based judgment. In line with previous research suggesting that forgetting does not act on abstracted knowledge like prototypes (Homa et al., 1973; Posner & Keele, 1970; Robbins et al., 1978) we found in Experiment 1 that participants in the linear judgment task were able to retain a high judgment accuracy even after a week, whereas participants in the multiplicative task made more errors in the delayed tests. In Experiment 2, however, a delayed test harmed judgment accuracy to the same degree in the multiplicative, exemplar-based judgment task as in the linear, rule-based judgment task. This result matches previous findings suggesting that actual rule-based judgments can also fluctuate over time (Balzer et al., 1983), but stands in contrast to research suggesting that abstracted knowledge is immune to forgetting (Homa et al., 1973; Posner & Keele, 1970; Robbins et al., 1978). One reason why people better retain rule-based judgments in Experiment 1 is possibly that they were able to apply the rules learned in training immediately to new test items so that the learned rules are less vulnerable to forgetting in the delayed tests. Limiting this opportunity to practice the judgment strategy, as in Experiment 2, may have restricted not only repeated retrieval of exemplars, but also the generalization of rules to new items. Taken together, those results point towards the view that not only exemplars may be forgotten over a longer time interval, but people may also experience difficulties to retrieve previously learned rules after a long time. Future research may seek to unravel on a more fine-grained level the degree to which specific mechanisms of forgetting, such as decay or interference, underlie forgetting in rule and exemplar retrieval.

When participants had to generalize the learned knowledge to new items, we found that participants in the multiplicative task were rather bad at judging new items. A plausible reason for this high number of judgment errors is that we selected the new items so that they strongly discriminate between the strategies, but both strategies did not generate a high performance on new items in the multiplicative task. Accordingly, we used the new items primarily to distinguish the judgment strategies and not to evaluate performance.

The stability of judgment strategies and exemplar memory over time

The question of how stable people's judgment strategies are over time is of high practical relevance (Ashton, 2000). One line of research has argued that people's judgment weights may fluctuate only to a small degree over time (Balzer et al., 1983), whereas other researchers have proposed that the time that has passed critically influences the strategy people follow (Bourne et al., 2006). Our study unites those divergent ideas: If participants had the opportunity to repeatedly practice their judgment strategy, we found that their judgment policies were highly consistent across time indicating that repeated practice can render people's judgment policies temporally more stable. However, if participants did not engage in exemplar retrieval for a long time as in Experiment 2, they shifted more towards rule-based strategies. These findings are consistent with the idea that people only engage in exemplar retrieval after a long time if the exemplars can still be retrieved. However, if previously encountered exemplars can no longer be retrieved, participants may revert to a less appropriate rule-based strategy (Bourne et al., 2006; Olsson et al., 2006), potentially inferring a linear relationship from any knowledge they can still recover. For instance, if people still remember which cues to focus their attention on or are able to retrieve at least two exemplars, they might use this knowledge to infer a linear, additive relationship.

To assess to what extent people still possess some memory for specific exemplars after a week, we additionally measured recognition memory in Experiment 2. Interestingly, the ability to discriminate old from new items varied in both judgment tasks as a function of the retention interval and strategy used: In the multiplicative task, exemplar and rule users were equally successful in discriminating between old and new items in immediate test; however, a week later, participants classified to the exemplar model more accurately discriminated between old and new items than participants classified to the rule-based model —a finding further supporting the idea that those participants who have a worse memory for previously encountered exemplars try to reinstate their judgment by applying rules. In turn, in the linear task, participants who were best described by the task-appropriate rule-based strategy better recognized old items than participants best described by the exemplar model. Furthermore, on average, participants more accurately discriminated between old and new items in the linear than the multiplicative task. This finding matches well with the idea that judgment accuracy potentially decreases in the multiplicative task because participants can no longer discriminate old from new items. In combination, our results highlight that rule-based learners likewise store a memory trace of previously seen exemplars (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004).

Restrictions in training duration

In our study, we tried to equate learning performance by setting a strict learning criterion, but participants could achieve this learning criterion after a variable number of learning blocks. We used this learning criterion because participants solving a multiplicative task often achieve a higher performance than participants in the linear task after the same number of training blocks (Hoffmann et al., 2014, 2016). In our study, participants in the multiplicative task also reached the learning criterion slightly faster than participants in the linear task. This result may hint at the alternative interpretation that training may have prematurely stopped in the multiplicative task and therefore participants may have remembered their judgments less well. Two arguments speak against this interpretation. First, if participants in the multiplicative task reached the learning criterion by chance, they should make more errors in the blocks preceding the last training block than participants who passed the learning criterion in the linear task. Yet, in the three blocks before the learning criterion was reached, judgment error is comparable in the linear and the multiplicative judgment task in most conditions. Second, if participants remembered their judgments less well because they solved a fewer number of training blocks, participants who needed more training blocks to reach the learning criterion should show a lower rate of forgetting. Yet, participants who reached the learning criterion in the multiplicative task after 15 or more training blocks made on average more errors than participants who reached the criterion earlier in training. Furthermore, judgment error increased more strongly from training to test for participants who needed more blocks. In sum, those results make it unlikely that training stopped too early in the multiplicative task.

Implications for training

From a broader perspective, considering which knowledge people are more likely to forget may inform our understanding about how people can best acquire this knowledge and retain it for a long time. For instance, the knowledge about categories that people retain after a longer time interval depends on how they learned the task (Sakamoto & Love, 2010). Yet people do not always structure their learning in a way that facilitates later retrieval, neither in education (Bjork, Dunlosky, & Kornell, 2013) nor when learning abstract concepts (Tauber, Dunlosky, Rawson, Wahlheim, & Jacoby, 2013). Our study contributes to a new branch of research in function learning and categorization studying how to construct specific training procedures to improve categorization decisions over a long time interval. This line of research has investigated how manipulations that improve long-term retention may help category or function learning and generalization, ranging from spaced training (Carvalho & Goldstone, 2014; McDaniel, Fadler, & Pashler, 2013; Zulkiply & Burt, 2013) to testing effects (Kang et al., 2011) to optimal training exemplars (Giguere & Love, 2013; Hornsby & Love, 2014). For instance, spacing exemplar presentations improves memory performance for trained items and simplifies generalization to new items (McDaniel et al., 2013). Our study emphasizes that identifying the underlying task structure and the strategies people use to approach the task can help to adapt those training procedures. Specifically, if rules can be abstracted as in linear judgment tasks, it may be sufficient to test those rules out on new items and distribute training across time to achieve high judgment accuracy and adequate generalization. In contrast, multiplicative tasks require that participants identify and retrieve specific exemplars. Introducing new probes interferes with retrieval of those exemplars suggesting that successful training procedures potentially need to tackle this identification problem.

Conclusions

Since Ebbinghaus's (1885) seminal work much research has been devoted to the study of forgetting. Our study highlights that forgetting prior knowledge can similarly restrict how accurately people make judgments after some time has passed —not only if people need to retrieve past experiences, but also if they need to established a judgment policy based on abstracted knowledge. Identifying how abstracted knowledge and past experiences can best be retained may thus help improve human judgments in different domains from weather forecasts to business.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. doi:10.1037/0278-7393.20.5.1063

Anderson, M. C. & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (2nd ed., pp. 237–313). San Diego, CA, US: Elsevier. doi:10.1016/B978-012102570-0/50010-0

Ashton, R. H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making*, *13*(3), 277–294. doi:10.1002/1099-0771(200007/09)13:3<277::AID-BDM350>3.0.CO;2-B

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, *113*(1), 1–29. doi:10.1037/0096-3445.113.1.30

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*(1), 54–75. doi:10.1037/0096-3445.104.1.54

Balzer, W. K., Rohrbaugh, J., & Murphy, K. R. (1983). Reliability of actual and predicted judgments across time. *Organizational Behavior & Human Performance*, *32*(1), 109–123. doi:10.1016/0030-5073(83)90142-3

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. doi:10.1146/annurev-psych-113011-143823

Bourne, L. E., Healy, A. F., Kole, J. A., & Graham, S. M. (2006). Strategy shifts in classification skill acquisition: Does memory retrieval dominate rule use? *Memory & Cognition*, *34*(4), 903–913. doi:10.3758/BF03193436

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*(2-3), 137–154. doi:10.1016/0001-6918(94)90048-5

Capaldi, E. J. & Neath, I. (1995). Remembering and forgetting as context discrimination. *Learning & Memory*, *2*(3-4), 107–132. doi:10.1101/lm.2.3-4.107

Carvalho, P. F. & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. doi:10.3758/s13421-013-0371-0

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods and applications*. San Diego, CA: Academic Press.

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. doi:10.1093/cercor/bhr036

Donkin, C., Newell, B. R., Kalish, M. L., Dunn, J. C., & Nosofsky, R. M. (2014). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933–948. doi:10.1037/xlm0000083

Ebbinghaus, H. (1885). *Über das Gedächtnis*. Leipzig: Duncker und Humblot.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465–485. doi:10.1037//0033-295X.86.5.465

Erickson, M. A. & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. doi:10.1037/0096-3445.127.2.107

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *115*(2), 155–174. doi:10.1037/0096-3445.115.2.155

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. doi:10.1016/S1364-6613(99)01294-2

Giguere, G. & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, *110*(19), 7613–7618. doi:10.1073/pnas.1219674110

Hahn, U. & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*(2-3), 197–230. doi:10.1016/S0010-0277(97)00044-9

Hasselmo, M. E. (2017). Avoiding catastrophic forgetting. *Trends in Cognitive Sciences*, *21*(6), 407–408. doi:10.1016/j.tics.2017.04.001

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242–2261. doi:10.1037/a0037989

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/xlm0000241

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116–122. doi:10.1037/h0035772

Hornsby, A. N. & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72–76. doi:10.1016/j.jarmac.2014.04.009

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 259–298. doi:10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156. doi:10.1037/0096-3445.132.1.133

Kang, S. H. K., McDaniel, M. A. M., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998–1005. doi:10.3758/s13423-011-0113-x

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic bulletin & review*, *14*(6), 1140–1146. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18229487

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making*, *3*(3), 244–260. Retrieved from http://journal.sjdm.org/bn5/bn5.html

Karpicke, J. D. & Roediger III, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. doi:10.1126/science.1152408

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. doi:10.1037/0033-295X.99.1.22

Little, J. L. & McDaniel, M. A. M. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*(2), 283–297. doi:10.3758/s13421-014-0475-1

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188–1201. doi:10.1037/a0026084

McCloskey, M. & Cohen, N. J. (1986). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, *24*, 109–165. doi:10.1016/S0079-7421(08)60536-8

McDaniel, M. A. M., Fadler, C. L. C., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1417–1432. doi:10.1037/a0032184

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*(4), 352–370. doi:10.1037/h0069819

Meeter, M., Murre, J. M. J., & Janssen, S. M. J. (2005). Remembering the news: Modeling retention data from a study with 14,000 participants. *Memory & Cognition*, *33*(5), 793–810. doi:10.3758/BF03193075

Morris, S. B. & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi:10.1037/1082-989X.7.1.105

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. doi:10.1037//0278-7393.14.4.700

Nosofsky, R. M. & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*(3), 375–402.

Nosofsky, R. M. & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesics and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*, 247–255. doi:10.1111/1467-9280.00051

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1371–1384. doi:10.1037/0278-7393.32.6.1371

Pachur, T. & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. doi:10.1016/j.cogpsych.2012.03.003

Palmeri, T. J. & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568. doi:10.1037/0278-7393.21.3.548

Platzer, C. & Bröder, A. (2013). When the rule is ruled out: Exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*(5), 429–441. doi:10.1002/bdm.1776

Posner, M. I. & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2), 304–308.

Postman, L. (1971). Organization and interference. *Psychological Review*, *78*(4), 290–302. doi:10.1037/h0031031

Richardson, D. C. & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295. doi:10.1016/S0010-0277(00)00084-6

Robbins, D., Barresi, J., Compton, P., Furst, A., Russo, M., & Smith, M. A. (1978). The genesis and use of exemplar vs. prototype knowledge in abstract category learning. *Memory & Cognition*, *6*(4), 473–480. doi:10.3758/BF03197481

Roediger III, H. L., Weinstein, Y., & Agarwal, P. K. (2010). Forgetting. In S. Della Sala (Ed.), *Forgetting: Preliminary considerations* (pp. 1–22). Hove, UK: Psychology Press. doi:10.4324/9780203851647

Roediger, H. L. & Karpicke, J. D. (2006). The Power of Testing Memory . *Perspectives on Psychological Science*, *1*(3), 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Rouder, J. N. & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63–82. doi:10.1037/0096-3445.133.1.63

Rubin, D. C. & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760. doi:10.1037/0033-295X.103.4.734

Sakamoto, Y. & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553. doi:10.1037/0096-3445.133.4.534

Sakamoto, Y. & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied*, *16*(4), 361–377. doi:10.1037/a0021610

Scholz, A., von Helversen, B., & Rieskamp, J. (2015). Eye movements reveal memory processes during similarity- and rule-based decision making. *Cognition*, *136*, 228–246. doi:10.1016/j.cognition.2014.11.019

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Siegel, J. A. & Siegel, W. (1972). Absolute judgment and paired-associate learning: Kissing cousins or identical twins? *Psychological Review*, *79*(4), 300–316. doi:10.1037/h0032945

Smithson, M. & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. doi:10.1037/1082-989X.11.1.54

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, *20*(2), 356–63. doi:10.3758/s13423-012-0319-6

von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology*, *46*(1), 867–889. doi:10.1037/a0016690

von Helversen, B. & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*(1), 73–96. doi:10.1037/0096-3445.137.1.73

von Helversen, B. & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 867–889. doi:10.1037/a0015501

Wagenmakers, E.-J. & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. doi:10.3758/BF03206482

Zulkiply, N. & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16–27. doi:10.3758/s13421-012-0238-9

Table 1

Task Structure in Experiment 1

Cue values				Linear Task			Multiplicative Task			Item Type
C_1	C_2	C_3	C_4	Criterion	Rule	Exemplar	Criterion	Rule	Exemplar	
0	0	0	0	10	10	10	10	9.1	10	Old
0	0	0	1	11	11	11	10	10.1	10	Old
0	0	1	0	12	12	12	11	9.9	11	Old
0	0	1	1	13	13	13	11	11.0	11	Old
0	1	0	1	14	14	14	12	12.7	12	Old
0	1	1	0	15	15	15	12	12.6	12	Old
0	1	1	1	16	16	16	13	13.6	13	Old
1	0	1	0	16	16	16	13	14.9	13	Old
1	1	1	0	19	19	19	18	17.5	18	Old
1	1	1	1	20	20	20	20	18.6	20	Old
0	1	0	0	13	13	13	11	11.7	11.3	New
1	0	0	0	14	14	13	12	14.1	11.5	New
1	0	0	1	15	15	11	12	15.1	10	New
1	0	1	1	17	17	16.3	14	15.9	14.7	New
1	1	0	0	17	17	19	14	16.7	18	New
1	1	0	1	18	18	17	16	17.8	16	New

Note. The judgment criterion was derived from Equation 1 (linear) and Equation 2 (multiplicative).

Table 2

Performance and Evidence Ratio in Experiment 1. Standard Error in Parentheses.

	Judgment Task					
	Linear			Multiplicative		
Training session						
Number of blocks		15.3 (0.7)			13.8 (0.7)	
Error last block		1.24 (0.17)			1.02 (0.16)	
Retention interval	Immediate	1 day	1 week	Immediate	1 day	1 week
Test session						
Error old items	1.16 (0.15)	1.20 (0.14)	1.19 (0.14)	1.29 (0.15)	1.45 (0.16)	1.55 (0.16)
Error new items	1.56 (0.15)	1.69 (0.18)	1.77 (0.19)	2.50 (0.17)	2.54 (0.17)	2.68 (0.18)
Evidence ratio	.28 (.07)	.31 (.07)	.12 (.05)	.53 (.08)	.48 (.08)	.36 (.08)

Note. Error was measured as the RMSD (Root Mean Squared Deviation) between participant's judgment and the criterion. The evidence ratio expresses how much support is provided for the exemplar strategy over the rule-based strategy ranging from 0 (evidence in favor of the rule-based strategy) to 1 (evidence in favor of the exemplar strategy).

Table 3

Performance and Evidence Ratio in Experiment 2. Standard Error in Parentheses.

	Judgment Task			
	Linear		Multiplicative	
	Retention interval		Retention interval	
	Immediate	1 week	Immediate	1 week
Training session				
Number of blocks	15.9 (0.7)	16.9 (0.6)	14.6 (0.7)	15.7 (0.7)
Error last block	1.33 (0.16)	1.52 (0.21)	1.34 (0.21)	1.63 (0.23)
Test session				
Error old items	1.24 (0.16)	1.73 (0.19)	1.49 (0.16)	2.37 (0.18)
Error new items	1.86 (0.19)	1.96 (0.15)	2.42 (0.14)	2.79 (0.16)
Evidence ratio	.23 (.07)	.14 (.05)	.69 (.08)	.15 (.06)
Recognition % correct	61.6 (3.4)	69.2 (2.9)	62.6 (3.1)	58.4 (2.9)

Note. Error was measured as the RMSD (Root Mean Squared Deviation) between participant's judgment and the criterion.

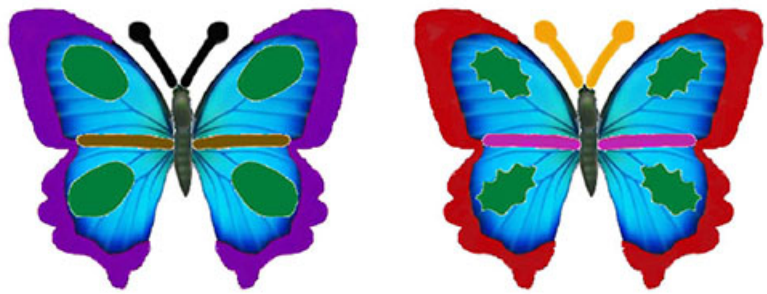

Figure 1. Sample species of butterflies with distinct cue values on all cues.

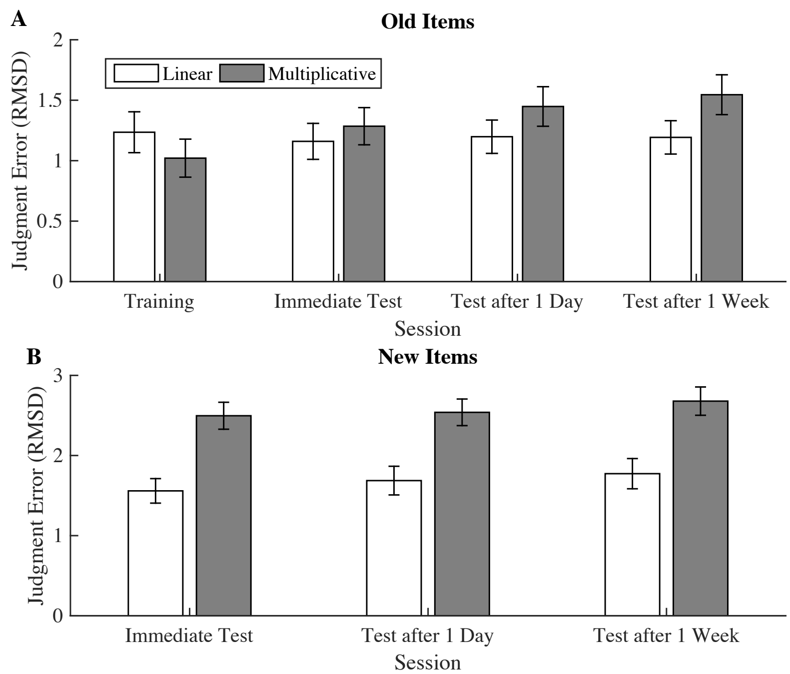

Figure 2. Judgment error measured in root-mean squared deviation (RMSD) on old items (Panel A) and new items (Panel B) in Experiment 1. White bars depict judgment error in the linear judgment task, gray bars depict judgment error in the multiplicative judgment task. **A.** Judgment error on old items was assessed for each participant in the last block of training as well as in all three test sessions (immediate test, test after 1 day, test after 1 week). **B.** Judgment error on new items was assessed for each participant in all three test sessions. Error bars indicate ± 1 *SE*.

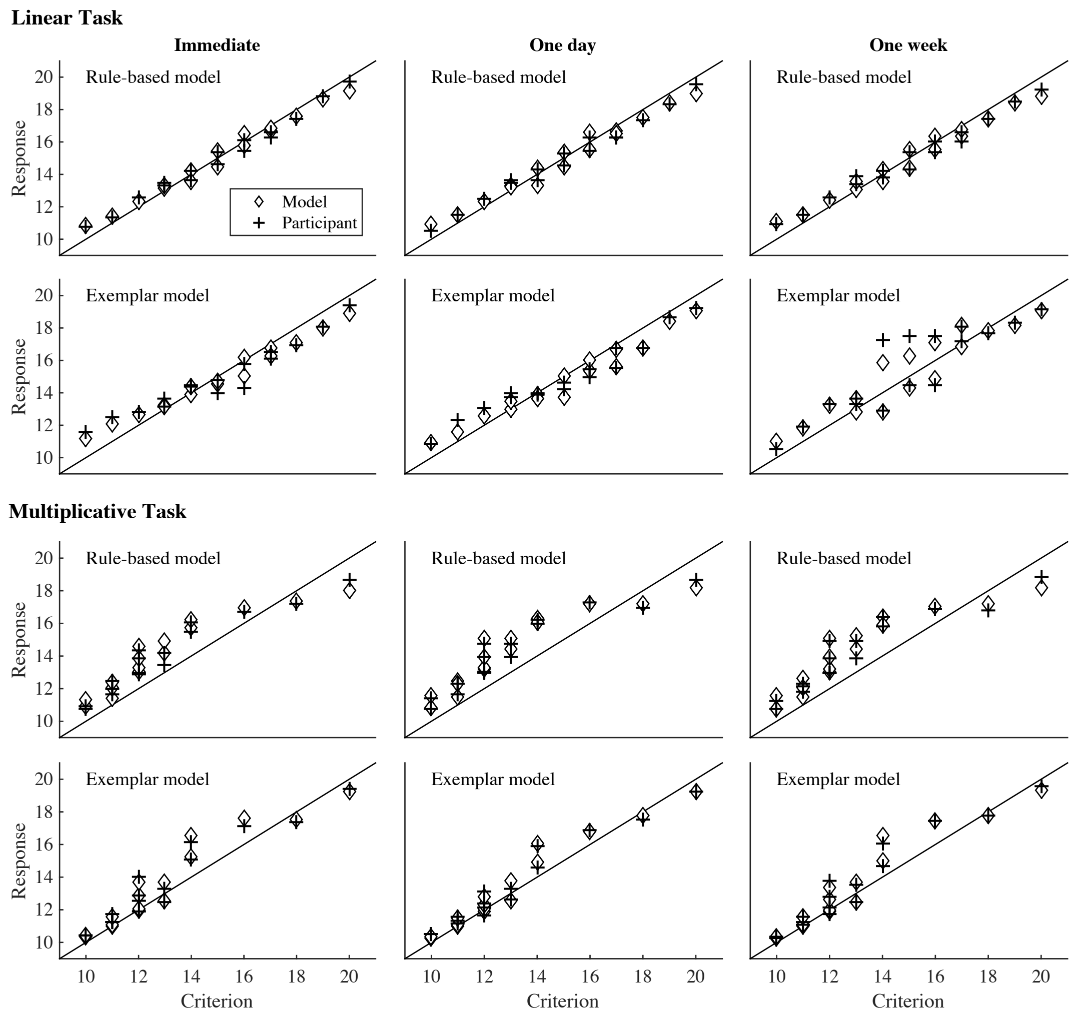

Figure 3. Model predictions and participants' judgments in Experiment 1 averaged across those participants clearly best described by either the rule-based model or the exemplar model, separately for the linear (upper two rows) and the multiplicative judgment task (lower two rows) and test session (immediate test, test after one day, or test after one week, in columns). Diamonds depict average model predictions; crosses depict participants' average judgments. The black diagonal lines depict perfectly accurate judgments.

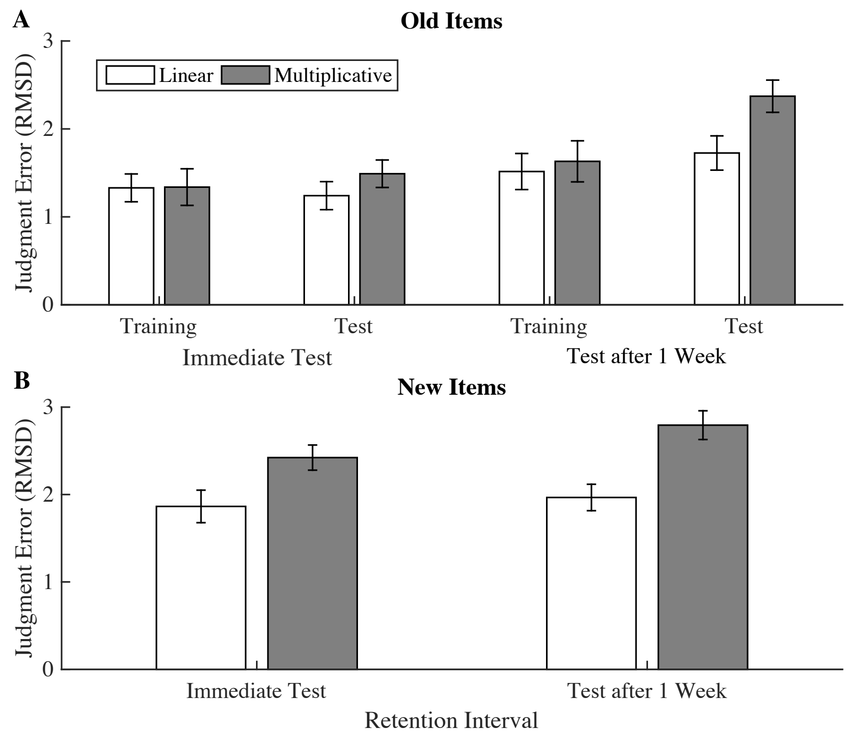

Figure 4. Judgment error measured in root-mean squared deviation (RMSD) on old items (Panel A) and new items (Panel B) in Experiment 2. White bars depict judgment error in the linear judgment task, gray bars depict judgment error in the multiplicative judgment task. **A.** Judgment error on old items was assessed for each participant in the last block of training as well as after a short retention interval (immediate test) or a long retention interval (test after one week). **B.** Judgment error on new items was assessed for each participant after either a short retention interval (immediate test) or a long retention interval (test after one week). Error bars indicate ± 1 *SE*

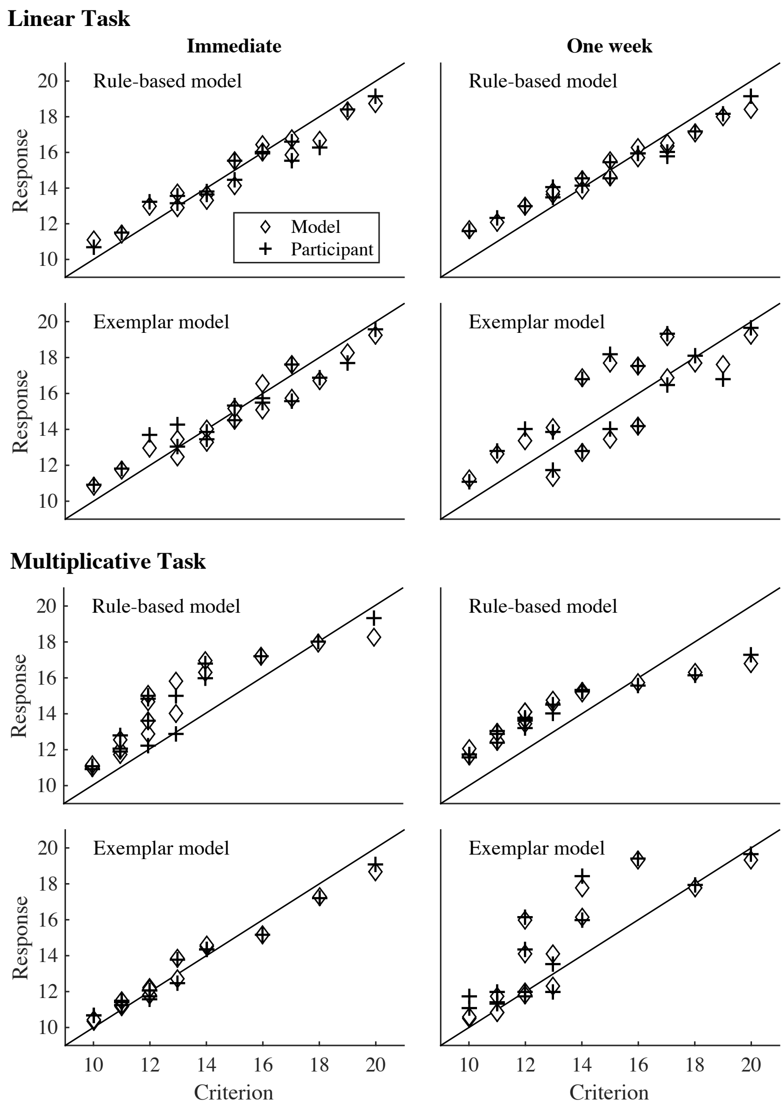

Figure 5. Model predictions and participants' judgments in Experiment 2 averaged across those participants clearly best described by either the rule-based model or the exemplar model, separately for the linear (upper two rows) and the multiplicative judgment task (lower two rows) and test session (immediate test or test after one week, in columns). Diamonds depict average model predictions; crosses depict participants' average judgments. The black diagonal lines depict perfectly accurate judgments.

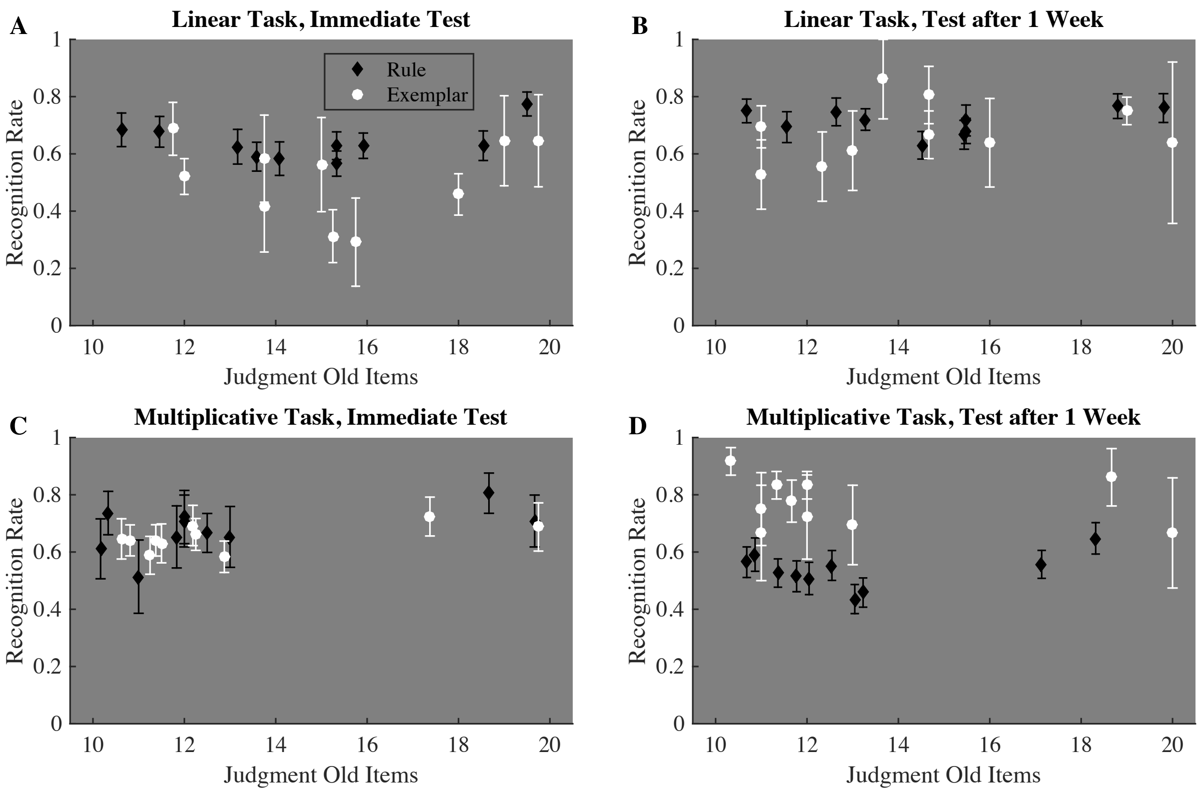

Figure 6. Proportion of correctly recognized old items (recognition rate) plotted against the average judgment for this old item in the last block of training, separately for participants classified to the rule-based (black diamonds) and the exemplar-based judgment strategy (white circles). Panel **A** depicts the recognition rate for participants who solved the linear task and took an immediate test. Panel **B** depicts the recognition rate for participants who solved the linear task and took the test after a week. Panel **C** and **D** show the recognition rate in the multiplicative task for immediate test and test after a week, respectively. Error bars indicate ± 1 *SE*.

.

Appendix A

Cognitive modeling of judgment strategies

To identify the cognitive strategies that people rely on in the three test sessions, we used a computational modeling approach. We compared how well a prominent rule-based model, a regression model, described participants' judgments in comparison to one often-used exemplar model using four attention weights. We compared all models to a guessing model that assumed that participants' judgments vary around participants' mean judgment on each trial. The guessing model estimated two free parameters: participants' mean judgment and the fitted dispersion parameter ϕ (see the paragraph on model estimation).

Model description

Rule-based model. To model rule-based strategies, we fitted a linear regression model that has often served as the prototypical rule-based strategy in judgment tasks (Cooksey, 1996; Juslin et al., 2003). The linear regression model allows combining several cues in a linear additive fashion. Accordingly, the estimated criterion value $\hat{y}_p$ of an object p is the weighted sum of the cue values x_{pi}:

$$\hat{y}_p = k + \sum_{i=1}^{4} w_i \cdot x_{pi} \tag{3}$$

where w_i are the cue weights for each cue i and k is a constant intercept. In sum, the linear model estimates six parameters: four cue weights w_i, one intercept k, and the dispersion parameter.

Exemplar model. Exemplar models have been widely used in judgment and categorization research to model retrieval of single instances from long-term memory (Hoffmann et al., 2014; Juslin et al., 2003). In exemplar models, the similarity $S(p, q)$ between the probe p and exemplar q is an exponential decay function of the distances d_{pq} between the objects (Nosofsky & Zaki, 1998).

$$S(p, q) = e^{-d_{pq}} \tag{4}$$

Thus, smaller distances between the probe p and exemplar q indicate a higher similarity between theses objects. To determine this distance, the cue values x_{pi} of probe p are compared to the cue values x_{qi} of exemplar q on all cues i. The more the cue values match each other, the smaller is the distance between the objects (Nosofsky & Johansen, 2000).

$$d_{pq} = h(\sum_{i=1}^{4} w_i|x_{pi} - x_{pq}|) \tag{5}$$

The sensitivity parameter h determines how strongly similarity decays with distance. Smaller sensitivity parameters indicate that similarity declines less with distance. The attention weights w_i, summing to one, weigh how much attention each cue or dimension receives. To account for judgments, Juslin et al. (2003) assumed that the criterion value c_q of an exemplar is stored together with its cue values in memory. To estimate the criterion value of a new probe $\hat{y}_p$, the criterion values y_q for each exemplar are weighted by the similarities.

$$\hat{y}_p = \frac{\sum_{q=1}^{Q} S(p,q) \cdot c_q}{\sum_{q=1}^{Q} S(p,q)} \tag{6}$$

In sum, the exemplar model estimates five parameters: three attention weights w_i, the sensitivity parameter h, and the dispersion parameter.

Model estimation and comparison. To evaluate the models' relative performance we fitted all models to participants' judgment on all six presentations of old and new items, separately for each of the three test sessions (immediate test, test after a day, and test after a week). The models were evaluated based upon the Bayesian Information Criterion (BIC; Schwarz, 1978). All models were fitted to participants' responses by minimizing the deviance -2LL, the negative summed log-likelihood L of the model given the data.

$$-2LL = -2 \cdot \sum \ln(L) \tag{7}$$

We calculated the likelihood for participants' judgments j assuming a beta distribution to account for the bounded scale between 10 and 20. Because the beta distribution is not defined for the open interval] 0, 1 [, participants' judgments and the correct criterion values for the exemplar model were re-scaled to the range 0.05 to 0.95. To make the interpretation of the parameter values easier, we followed the method for the beta regression suggested by Smithson and Verkuilen (2006) and re-formulated the shape parameters so that they represent a location parameter μ and a dispersion parameter ϕ (for a detailed explanation see Smithson & Verkuilen, 2006). The location parameter μ represented the model's prediction for this item, $\hat{y_p}$, whereas the dispersion parameter ϕ was estimated based on $\phi = \mathrm{e}^{-b_0}$. Because the rule-based model would allow to predict criterion values below 0 or above 1, we used a logistic link function as in the logistic regression to strictly keep the model's predictions in range. The estimated parameter values in the beta regression can then be interpreted in a similar fashion as in the logistic regression. For the rule-based model we provide here all estimated location parameters in the form of odds ratios (as in the beta regression analyses on evidence ratios), whereas the dispersion parameter is not transformed. Note that we do not need to apply the logistic link to the location parameter of the exemplar model because it does not extrapolate beyond the range of encountered values. Accordingly, the estimated parameter values in the exemplar model do not change their meaning. The likelihood of the beta distribution can finally be formulated as

$$L = Beta(\hat{y_p}\phi, \phi - \hat{y_p}\phi; j) \tag{8}$$

To compare which model described participants' responses better, we calculated the BIC for each model. The BIC can be used to compare non-nested models and penalizes more complex models by accounting for the number of free model parameters k:

$$\mathrm{BIC} = -2LL + k \ln n, \tag{9}$$

where n denotes the number of observations. Smaller BIC values indicate a better model fit. BICs were converted into BIC weights $\mathrm{BIC}_{w,M}$ that give the posterior probability of each model given the data (Wagenmakers & Farrell, 2004).

$$\mathrm{BIC}w_M = \frac{e^{-.5\Delta\mathrm{BIC}_M}}{\sum_i e^{-.5\Delta\mathrm{BIC}_i}} \tag{10}$$

with Δ BIC$_M$ as the difference between model M and the best model MB in the set and Δ BIC$_i$ as the difference between the model i and the best model.

Detailed results for model comparisons in Experiment 1

Table A1 lists the mean and standard deviation of the model parameters as well as b_0. To facilitate interpretability, the weights w_i and the intercept k in the rule model as well as the guessing model were transformed into odds ratios. Attention weights w_i do not need to be transformed into odds ratios, but were transformed to reflect relative attention weights and the sensitivity parameter h.

Table A2 shows the average BIC for each model, average BIC weights, and strategy classifications for each task and retention interval. Overall, the average BIC was lower for the rule-based model in all tasks and conditions. Yet, to what degree the BIC favoured one over the other model strongly varied across participants, and ranged for instance in the immediate test session from $\Delta\mathrm{BIC}_{Rule-Ex}$ = -3167 to $\Delta\mathrm{BIC}_{Rule-Ex}$ = 120. Furthermore, if we only consider $\Delta\mathrm{BIC}_{M-MB} < -5$ as sufficient evidence for one model (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2014), not all participants were consistently classified to one model in the linear or the multiplicative task (Table A2) and it varied across sessions for which participant a strategy could be clearly identified. Because BIC differences and classification consistency vary across participants, we quantified the evidence in favour of each model using BIC weights. On average, these BIC weights provide more evidence for the rule-based model in the linear task, but evidence is more evenly distributed among the rule-based and the exemplar-based model in the multiplicative task. When finally analyzing strategy use over time, we excluded those participants for whom the evidence for the guessing model outweighed the evidence for the rule-based and the exemplar model ($BIC_w > .5$) in any session (three participants in both tasks) and calculated the evidence ratio for the exemplar model over the rule-based model as a normalized probability, $\text{BIC}_{w,Ex}/(\text{BIC}_{w,Ex} + \text{BIC}_{w,Rule})$.

Detailed results for model comparisons in Experiment 2

Table A4 displays BIC weights, strategy classifications based on BIC weights, and the RMSD between model responses and participants' judgments. Similar to Experiment 1, BIC weights for the guessing model as well as the number of participants classified to the guessing model were low with a slightly higher number of participants classified to the guessing model in the multiplicative task. In the linear task, the majority of participants were best described by the rule-based model both in the immediate test session as well as in test after one week, as shown by high average BIC weights for the rule-based model and a high number of participants classified to that model. The exemplar model better described participants' judgment in the immediate test in the multiplicative task, as suggested by higher BIC weights and more participants classified to the model. In test after one week, however, the rule-based model provided a higher BIC weight and more participants were classified to the rule-based model.

Table A3 lists the mean and standard deviation of the model parameters; table A4 shows the average BIC for each model, average BIC weights, and strategy classifications for each participant. As in Experiment 1, the average BIC is in most conditions lower for the rule-based model than for the exemplar model, except for immediate test in the multiplicative task. Similarly, the BIC weights provide a higher evidence for the rule-based model than for the exemplar model or the guessing model in most conditions, but provide a higher evidence for the exemplar model in immediate test in the multiplicative task. As in Experiment 1, not all participants were consistently classified to one strategy based upon the differences in BIC. Furthermore, slightly more participants were classified to a guessing model in the multiplicative than in the linear task. As in Experiment 1, we excluded those participants for whom the evidence for the guessing model outweighed the evidence for the rule-based and the exemplar model and calculated the evidence for a rule-based model over an exemplar model for each of the remaining participants.

Table A1

Model parameters in Experiment 1. SD in Paranthesis

RI	Task	Model	w_1	w_2	w_3	w_4	k	h	b_0
Imm.	Lin	Guess	–	–	–	–	1.0 (0.1)	–	-1.1 (0.3)
		Rule	4.1 (2.3)	3.5 (1.9)	2.5 (1.7)	1.5 (0.5)	0.2 (0.2)	–	-2.9 (1.0)
		Ex	.29 (.20)	.29 (.31)	.28 (.24)	.15 (.28)	–	85 (71)	-2.6 (0.8)
	Mult	Guess	–	–	–	–	0.7 (0.2)	–	-0.9 (0.3)
		Rule	6.8 (5.0)	3.3 (2.1)	1.8 (1.1)	1.6 (0.7)	0.2 (0.2)	–	-3.0 (4.8)
		Ex	.40 (.29)	.15 (.19)	.15 (.19)	.30 (.36)	–	82 (75)	-2.3 (0.8)
1 day	Lin	Guess	–	–	–	–	1.0 (0.1)	–	-1.1 (0.3)
		Rule	4.6 (3.5)	3.6 (2.2)	2.4 (1.1)	1.5 (0.6)	0.2 (0.2)	–	-3.1 (1.2)
		Ex	.32 (.21)	.25 (.25)	.30 (.21)	.13 (.24)	–	98 (77)	-2.7 (0.8)
	Mult	Guess	–	–	–	–	0.7 (0.2)	–	-0.9 (0.3)
		Rule	6.9 (5.2)	3.6 (2.7)	2.2 (1.6)	1.9 (1.8)	0.1 (0.2)	–	-2.6 (0.8)
		Ex	.52 (.31)	.12 (.19)	.25 (.29)	.11 (.26)	–	89 (59)	-2.4 (0.8)
1 week	Lin	Guess	–	–	–	–	1.0 (0.2)	–	-1.1 (0.3)
		Rule	4.5 (3.3)	4.1 (3.7)	2.5 (1.6)	1.5 (0.7)	0.2 (0.2)	–	-3.2 (1.2)
		Ex	.38 (.22)	.22 (.24)	.30 (.22)	.09 (.16)	–	106 (81)	-2.7 (0.85)
	Mult	Guess	–	–	–	–	0.7 (0.2)	–	-0.9 (0.3)
		Rule	7.4 (5.5)	3.4 (2.3)	2.6 (3.7)	1.8 (1.3)	0.1 (0.2)	–	-2.6 (0.7)
		Ex	.42 (.31)	.16 (.24)	.18 (.23)	.23 (.35)	–	107 (76)	-2.4 (0.8)

Note. RI = Retention interval, Imm. = Immediate, Guess = Guessing model, Rule = Rule-based model, Ex = Exemplar model, w_1 - w_4 = cue weights in the linear model and attention weights in the exemplar model, respectively, k = Intercept in the linear model, h = sensitivity parameter in the exemplar model, b_0 = dispersion.

Table A2

BIC, BIC_w, and strategy classification in Experiment 1. SD for BIC weights in Paranthesis

Indicator	Model	Linear Task			Multiplicative Task		
		Immediate	One day	One Week	Immediate	One day	One Week
BIC	Guess	-5 (17)	-4 (15)	-3 (10)	-9 (21)	-8 (15)	-8 (16)
	Rule	-157 (96)	-175 (111)	-186 (107)	-188 (484)	-140 (74)	-143 (69)
	Exemplar	-129 (69)	-140 (77)	-137 (81)	-106 (79)	-120 (87)	-120 (87)
BIC_w	Guess	.07 (.24)	.05 (.22)	.03 (.16)	.05 (.22)	.03 (.16)	.04 (.19)
	Rule	.67 (.42)	.67 (.45)	.84 (.35)	.43 (.48)	.50 (.49)	.60 (.48)
	Exemplar	.26 (.39)	.28 (.42)	.13 (.31)	.51 (.49)	.47 (.50)	.36 (.47)
N_{CL}	incons.	9	5	3	4	3	4
	Guess	2	2	1	2	1	1
	Rule	23	24	33	16	18	22
	Exemplar	6	9	3	18	18	13

Note. SD = Standard Deviation; BIC = Bayesian Information Criterion, BIC_w = Bayesian Information Criterion weights, N_{CL} = number of participants classified, incons. = inconsistent classification, Guess = Guessing model, Rule = Linear model, Ex = Exemplar model.

Table A3

Model parameters in Experiment 2. SD in Parenthesis

RI	Task	Model	w_1	w_2	w_3	w_4	k	h	b_0
Imm.	Lin	Guess	–	–	–	–	1.0 (0.2)	–	-1.1 (0.3)
		Rule	4.5 (3.1)	2.9 (1.5)	3.5 (4.5)	1.5 (0.6)	0.2 (0.3)	–	-3.0 (0.9)
		Ex	.38 (.25)	.14 (.20)	.37 (.28)	.11 (.22)	–	95 (73)	-2.7 (0.8)
	Mult	Guess	–	–	–	–	0.7 (0.2)	–	-0.9 (0.4)
		Rule	5.4 (4.7)	2.9 (1.8)	2.3 (1.9)	1.4 (0.8)	0.2 (0.2)	–	-2.0 (0.7)
		Ex	.33 (.26)	.24 (.26)	.19 (.23)	.24 (.31)	–	88 (70)	-2.2 (0.7)
1 week	Lin	Guess	–	–	–	–	1.1 (0.3)	–	-1.0 (0.3)
		Rule	4.1 (2.9)	3.3 (3.6)	2.3 (2.1)	1.5 (0.6)	0.5 (0.8)	–	-2.5 (1.0)
		Ex	.38 (.31)	.19 (.23)	.21 (.23)	.23 (.32)	–	104 (81)	-2.1 (0.85)
	Mult	Guess	–	–	–	–	0.8 (0.2)	–	-1.0 (0.6)
		Rule	3.9 (3.5)	2.7 (2.7)	1.7 (1.5)	1.4 (0.8)	0.3 (0.4)	–	-1.8 (0.6)
		Ex	.39 (.39)	.24 (.32)	.16 (.23)	.21 (.37)	–	64 (70)	-1.6 (0.6)

Note. RI = Retention interval, Guess = Guessing model, Rule = Rule-based model, Ex = Exemplar model, w_1 - w_4 = cue weights in the linear model and attention weights in the exemplar model, respectively, k = Intercept in the linear model, h = sensitivity parameter in the exemplar model, b_0 = dispersion.

Table A4

BIC, BIC_w, and strategy classification in Experiment 2. SD in Parenthesis

Indicator	Model	Linear Task		Multiplicative Task	
		Immediate	One Week	Immediate	One Week
BIC	Guess	-4 (15)	-3 (17)	-17 (30)	-20 (51)
	Rule	-164 (89)	-124 (85)	-96 (59)	-75 (53)
	Exemplar	-134 (74)	-85 (81)	-100 (77)	-46 (64)
BIC_w	Guess	.03 (.17)	.05 (.22)	.11 (.31)	.12 (.32)
	Rule	.75 (.40)	.81 (.35)	.27 (.40)	.74 (.41)
	Exemplar	.23 (.38)	.14 (.30)	.62 (.46)	.13 (.31)
N_{CL}	incons.	6	7	6	6
	Guess	1	1	3	4
	Rule	24	25	6	22
	Exemplar	4	3	16	3

Note. BIC = Bayesian Information Criterion, BIC_w = Bayesian Information Criterion weights, N_{CL} = number of participants classified, incons. = inconsistent classification, Guess = Guessing model, Rule = Linear model, Ex = Exemplar model.

Appendix B

Modeling forgetting of exemplars

We modelled forgetting of exemplars by using a successful exemplar-based learning model, ALCOVE (Kruschke, 1992). ALCOVE assumes that learning in judgment tasks can be understood as gradually forming associative links between the exemplars that are encountered and the possible criterion values. Judgments are a function of the similarity of the probe to the previously encountered exemplars and of the association strengths between the exemplars and the criterion values. That is, the probe activates similar exemplars, which in turn activate criterion values they are associated with. Association strengths are then translated into output probabilities for each criterion value and the final judgment is the mean of the criterion values weighted by their probabilities. ALCOVE contains three free parameters, two learning parameters and a sensitivity parameter: The first learning parameter determines the speed with which the associations between criterion values and exemplars are formed. The second learning parameter determines how fast people learn to differentially distribute attention to the features of the objects and changes how similarity between the probe and learnt exemplars is computed. The sensitivity parameter regulates how similarity is translated into the activation of an exemplar.

We introduced forgetting in ALCOVE by assuming that over time further exemplars would be encountered that interfere with previously learnt information. New information updates both the association between exemplars and stored criterion values as well as learned attention towards specific cues. For each simulation, we randomly drew 1000 times from an exponential distribution for the two learning parameters (with $M = 1$) and the sensitivity parameter (with $M = 3$). The value of 1 for the learning parameters was chosen to mimic the pattern that usually the values of the learning parameters are quite small, but may in judgment tasks also take values larger than 1. For the sensitivity parameter, we assumed that in our task participants mostly have a specific representation of the exemplars (most sensitivity parameter values are above 1), but some participants may have a more unspecific representation. In addition, we estimated ALCOVE's parameters for each individual in Experiment 1 using the training data and performed the same simulations with the bootstrapped parameters.

Training followed the same schedule as in the experiment, but we introduced four additional random cues in the simulation to limit catastrophic forgetting (French, 1999), a well-known problem in machine learning (Hasselmo, 2017; McCloskey & Cohen, 1986). Specifically, if we used only four cues and ALCOVE learned new random patterns, it would instantaneously forget everything it learned so far, as the new items that are described by the same cues as the old items would require the model to overwrite the learnt associations. To avoid this problem, a common strategy is to add a few more random cues, so that the new items that are learned do not completely overlap with the old information. To introduce forgetting, ALCOVE continued to learn N random item profiles between training and test (N varied from 0 to 100 in steps of 10). Finally, ALCOVE made the same judgments for old exemplars in test as participants did.